A Survey of Large Language Models

Abstract

Language is essentially a complex, intricate system of human expressions governed by grammatical rules. It poses a significant challenge to develop capable AI algorithms for comprehending and grasping a language. As a major approach, language modeling has been widely studied for language understanding and generation in the past two decades, evolving from statistical language models to neural language models. Recently, pre-trained language models (PLMs) have been proposed by pre-training Transformer models over large-scale corpora, showing strong capabilities in solving various NLP tasks. Since researchers have found that model scaling can lead to performance improvement, they further study the scaling effect by increasing the model size to an even larger size. Interestingly, when the parameter scale exceeds a certain level, these enlarged language models not only achieve a significant performance improvement but also show some special abilities that are not present in small-scale language models. To discriminate the difference in parameter scale, the research community has coined the term large language models (LLM) for the PLMs of significant size. Recently, the research on LLMs has been largely advanced by both academia and industry, and a remarkable progress is the launch of ChatGPT, which has attracted widespread attention from society. The technical evolution of LLMs has been making an important impact on the entire AI community, which would revolutionize the way how we develop and use AI algorithms. In this survey, we review the recent advances of LLMs by introducing the background, key findings, and mainstream techniques. In particular, we focus on four major aspects of LLMs, namely pre-training, adaptation tuning, utilization, and capacity evaluation. Besides, we also summarize the available resources for developing LLMs and discuss the remaining issues for future directions.